A set of squares without arithmetic progressions

by

Katalin Gyarmati and Imre Z. Ruzsa (Budapest)

To Andrzej Schinzel, with respect and gratitude

1. Introduction. The problem of finding arithmetic progressions in a partition of integers, or in a dense subset of the first $N$ integers, is among the oldest and most investigated questions of combinatorial number theory. We focus on the analogous problem for the first $N$ squares.

Let $Q(N)$ denote the maximal cardinality of sets $A \subset \{1^2, 2^2, \ldots, N^2\}$ which do not contain any nontrivial three-term arithmetic progression. The most fundamental question about this quantity, which we are unable to answer, is definitely the following.

Problem. Is $Q(N) = o(N)$?

We do not even have a convincing heuristic argument for one answer or the other. The only reason why we may be inclined to expect a positive answer is that so far we failed to construct such a set with positive density.

We are going to show that $Q(N)/N$ cannot tend to 0 too fast, which probably means that if it does so at all, this will be difficult to confirm.

Theorem. For every sufficiently large $N$ there is a set $A \subset \{1, \ldots, N\}$ such that the equation

$$x^2 + y^2 = 2z^2$$

has no solution with $x, y, z \in A$ other than the trivial solutions $x = y = z$, and

$$|A| > cN/\sqrt{\log \log N}$$

with a positive constant $c$.

We are slightly more confident about the partition version.
Conjecture. If we split the set of positive integers into finitely many parts, then the equation \( x^2 + y^2 = 2z^2 \) has a nontrivial solution with \( x, y, z \) being in the same part.

2. Proof. We call a solution of our favourite equation

\[
(2.1) \quad x^2 + y^2 = 2z^2
\]

primitive if \( x, y, z \) are coprime. Clearly every nonzero solution can be written as \( x = dx', y = dy', z = dz' \), where \( d = \gcd(x, y, z) \) and \( x', y', z' \) is a primitive solution. We will call this primitive solution \((x', y', z')\) the stem of the solution \((x, y, z)\).

Lemma 1. If \( x, y, z \) form a primitive solution of \((2.1)\), then \( x, y \) consist exclusively of primes \( p \equiv \pm 1 \pmod{8} \), and \( z \) consists exclusively of primes \( p \equiv 1 \pmod{4} \).

This reformulates the well-known property of the quadratic character of 2 and \(-1\).

For an integer \( j \), \( 1 \leq j \leq 7 \), let \( \nu_j(n) \) denote the number of prime divisors \( p \) of \( n \) satisfying \( p \equiv j \pmod{8} \), counted with multiplicity. These are completely additive functions.

Lemma 2. Let \( x, y, z \) be a solution of \((2.1)\). Write \( x = dx', y = dy', z = dz' \), where \( d = \gcd(x, y, z) \) and \((x', y', z')\) is its stem. We have

\[
(2.2) \quad \nu_5(x) - \nu_5(z) = -\nu_5(z'),
\]

\[
(2.3) \quad \nu_7(x) - \nu_7(z) = \nu_7(x').
\]

Proof. Indeed, \( \nu_5(x) = \nu_5(d) + \nu_5(x') = \nu_5(d) \) by the previous lemma and \( \nu_5(z) = \nu_5(d) + \nu_5(z') \); by subtracting we get \((2.2)\). Similarly \( \nu_7(x) = \nu_7(d) + \nu_7(x') \) and \( \nu_7(z) = \nu_7(d) + \nu_7(z') = \nu_7(d) \); by subtracting we get \((2.3)\). \( \square \)

Now we introduce the completely additive function

\[
\rho(n) = \nu_5(n) - \nu_7(n).
\]

Lemma 3. Let \( A \) be a set of integers with the property that \( \rho(n) = k \) for all \( n \in A \). Let \((x, y, z) \in A^3\) be a solution of \((2.1)\) with stem \((x', y', z')\). The three integers \( x', y', z' \) consist exclusively of primes \( p \equiv 1 \pmod{8} \).

Proof. By subtracting \((2.2)\) from \((2.3)\) we obtain

\[
\rho(z) - \rho(x) = \nu_7(x') + \nu_5(z').
\]

By the symmetric role of \( x \) and \( y \) we also have

\[
\rho(z) - \rho(y) = \nu_7(y') + \nu_5(z').
\]

On the left hand side of each equation we have 0 and on the right hand side a sum of nonnegative numbers, hence the numbers on the right hand side all
vanish. Since Lemma 1 already excludes the classes 3 and 5 (mod 8) for \( x' \) and \( y' \), as well as the classes 3 and 7 (mod 8) for \( z' \), only the class 1 (mod 8) remains.

By the Turán–Kubilius inequality we know that for most \( n \leq N \) the values of \( \rho(n) \) fall into an interval of length \( O(\sqrt{\log \log N}) \), so if we could exclude primitive solutions arising from primes in the congruence class 1 (mod 8) without much loss, we would be done. In what follows we achieve this.

**Lemma 4.** Let \( (x, y, z) \) be a primitive solution of (2.1) with \( x > z > y \). There are coprime positive integers \( u, v \) of opposite parity such that
\[
x = u^2 - v^2 + 2uv, \quad y = |u^2 - v^2 - 2uv|, \quad z = u^2 + v^2.
\]

**Proof.** By looking at the residues modulo 4 we see that \( x, y, z \) must all be odd. We can now rewrite equation (2.1) as
\[
\left( \frac{x + y}{2} \right)^2 + \left( \frac{x - y}{2} \right)^2 = z^2
\]
and apply the familiar parametric representation of Pythagorean triples. ■

Let \( W \subset \mathbb{N}^2 \) be the set of pairs \((u, v)\) which generate a triplet \((x, y, z)\) in the representation described in Lemma 4 such that \( x, y, z \) consist exclusively of primes \( p \equiv 1 \) (mod 8).

**Lemma 5.**
\[
|W \cap [1, N]^2| = O(N^2 (\log N)^{-3/2}).
\]

**Proof.** For a fixed value of \( u \) write
\[
W_u = \{ v : 1 \leq v \leq N, (u, v) \in W \}.
\]
First we estimate \( |W_u| \).

Let \( p \) be an odd prime, \( p \not\equiv 1, 3 \) (mod 8). We show that certain residue classes modulo \( p \) are missing from \( W_u \).

If \( p \mid u \), then the class of 0 is missing by coprimality and we cannot claim anything more.

Assume now \( p \nmid u, p \equiv 5 \) (mod 8). Let \( i \) be the solution of the congruence
\[
i^2 \equiv -1 \pmod{p}.
\]
The assumption that \( p \nmid z = u^2 + v^2 \) can be rewritten as
\[
v \not\equiv \pm iu \pmod{p},
\]
which yields two excluded residue classes.

Assume next \( p \nmid u, p \equiv 7 \) (mod 8). Let \( i \) be the solution of the congruence
\[
i^2 \equiv 2 \pmod{p}.
\]
The assumption that
\[ p \nmid x = u^2 - v^2 + 2uv = 2u^2 - (u - v)^2 \]
can be rewritten as
\[ v \not\equiv (\pm i + 1)u \pmod{p}, \]
which yields two excluded residue classes.

The assumption that
\[ p \nmid \pm y = u^2 - v^2 - 2uv = 2u^2 - (u + v)^2 \]
can be rewritten as
\[ v \not\equiv (\pm i - 1)u \pmod{p}, \]
and it yields another two excluded residue classes. It is easily seen that these four classes are distinct, so altogether we have four excluded classes.

By a familiar sieve estimate (e.g. Theorem 2.2 in Halberstam and Richert’s book [2]) we obtain
\[
|W_u| < c_1 N \prod_{p | u} \left(1 - \frac{1}{p}\right) \prod_{p | u, p \equiv 5 \pmod{8}, p < \sqrt{N}} \left(1 - \frac{2}{p}\right) \prod_{p | u, p \equiv 7 \pmod{8}, p < \sqrt{N}} \left(1 - \frac{4}{p}\right),
\]
where
\[
f(u) = \prod_{p | u, p \equiv 5 \pmod{8}} \frac{p - 1}{p - 2} \prod_{p | u, p \equiv 7 \pmod{8}} \frac{p - 1}{p - 4}.
\]

By using Dirichlet’s classical estimate
\[
\sum_{p \leq x, p \equiv j \pmod{8}} \frac{1}{p} = \frac{1}{4} \log \log x + O(1)
\]
for \( j = 5 \) and \( 7 \) we get
\[
|W_u| < c_2 f(u) N (\log N)^{-3/2}.
\]

Our function \( f(u) \) is unbounded, but it is bounded in mean:
\[
\sum_{u \leq N} f(u) < c_3 N.
\]

Estimates for sums of multiplicative functions that include the above one can be found in many places, for instance Corollary 5.1 in Tenenbaum’s book [6]. This implies the claim of the lemma.
Lemma 6.

\[ \sum_{(u,v) \in W} \frac{1}{u^2 + v^2} < \infty. \]

Proof. This follows from the previous lemma by partial summation. ■

Lemma 7. Let \( V \) be a set of positive integers and let \( B \) be the set of those positive integers that are not divisible by any element of \( V \). The set \( B \) has an asymptotic density and it is at least

\[ \prod_{v \in V} \left( 1 - \frac{1}{v} \right). \]

This is the Heilbronn–Rohrbach inequality (see e.g. [3]).

Proof of the Theorem. Let \( B \) be the set of integers which are not divisible by any number of the form \( u^2 + v^2 \), \((u,v) \in W\). By the previous lemma this set has a positive asymptotic density, say \( c_3 \). Now put

\[ A_k = \{ n \in B : n \leq N, \rho(n) = k \} \]

with a suitable \( k \). We claim that

(i) equation (2.1) has no nontrivial solution in any \( A_k \),
(ii) for a suitable \( k \) (depending on \( N \)) we have

\[ |A_k| > cN/\sqrt{\log \log N}. \]

These claims together clearly imply the Theorem.

For claim (i), suppose on the contrary that there is a solution \( x, y, z \) with stem \( x', y', z' \). By Lemma 3 these latter three integers consist only of primes \( \equiv 1 \pmod{8} \). Hence they are generated by some \((u,v) \in W\) and we would have

\[ u^2 + v^2 = z' \mid z \in A_k \subset B, \]

a contradiction with the definition of \( B \).

To show claim (ii), recall that the Turán–Kubilius inequality tells us

\[ \sum_{n=1}^{N} (\rho(n) - m)^2 < c_4 N \sum_{p^k \leq N} p^{-k} \rho(p^k)^2 < c_5 N \log \log N, \]

where

\[ m = \sum_{p \leq N} \rho(p)/p. \]

In particular, with a well-chosen \( c_6 \) there are \( < (c_3/2)N \) integers up to \( N \) such that

\[ |\rho(n) - m| \geq c_6 \sqrt{\log \log N}. \]
Omit these from \( B \); the rest still has \( > (c_3/2)N \) elements up to \( N \), and for some of the at most \( 2c_6\sqrt{\log \log N} \) possible values of \( \rho(n) \) at least one will appear \( cN/\sqrt{\log \log N} \) times. \( \square \)

3. Concluding remarks. Besides three-term progressions, characterized by the equation \( x + y = 2z \), one can consider the more general arithmetic-mean equation

\[
x_1 + \cdots + x_k = ky.
\]

Let \( Q_k(N) \) denote the maximal cardinality of sets \( A \subset \{1^2, 2^2, \ldots, N^2\} \) which do not contain any nontrivial solution of this equation (so that \( Q(N) = Q_2(N) \)). It is not difficult (though not quite obvious) to show \( Q_k(N) = o(N) \) for \( k \geq 6 \). Ben Green outlined to the authors a method that would prove this claim for \( k = 4 \), with the possibility of giving an effective estimate. This seems to be a limit to analytic methods.

It is not easy to estimate this quantity from below either. Let \( R_k(N) \) denote the maximal cardinality of sets \( A \subset [1, N] \) which do not contain any nontrivial solution of this equation. By a general theorem of Komlós, Sulyok and Szemerédi [4] (see also [5]) we know that \( Q_k(n) \gtrsim R_k(n) \). The best known lower estimate of \( R_k(N) \) is

\[
R_k(N) \gtrsim N \exp(-c_k \sqrt{\log N}),
\]

Behrend’s bound [1] with obvious changes. Can one do any better?

**Problem.** Is

\[
Q_3(N) \gtrsim N(\log N)^{-c}
\]

with some constant \( c \)?

While it is unlikely that the asymptotic behaviour of these quantities will be known in the near future, still it may be possible to compare them.

**Problem.** Given an integer \( k \geq 2 \), is there another integer \( l \) such that \( Q_l(N) \lesssim R_k(N) \)?

Acknowledgements. The authors were supported by ERC–AdG grant no. 228005 and Hungarian National Foundation for Scientific Research (OTKA) grants no. K67676, K72731, K81658 and PD72264.

References


A set of squares without arithmetic progressions


Katalin Gyarmati
Algebra and Number Theory Department
Eötvös Loránd University
H-1117 Budapest, Hungary
E-mail: gykati@cs.elte.hu

Imre Z. Ruzsa
Alfréd Rényi Institute of Mathematics
Pf. 127, H-1364 Budapest, Hungary
E-mail: ruzsa@renyi.hu

Received on 23.7.2011
and in revised form on 24.2.2012